

The Allosteric Switching Mechanism in Bacteriophage MS2

Matthew R. Perkett,^{1, a)} Dina T. Mirijanian,^{1, a)} and Michael F. Hagan^{1, b)}
Martin Fisher School of Physics, Brandeis University, Waltham, MA, USA

In this article we use all-atom simulations to elucidate the mechanisms underlying conformational switching and allostery within the coat protein of the bacteriophage MS2. Assembly of most icosahedral virus capsids requires that the capsid protein adopt different conformations at precise locations within the capsid. It has been shown that a 19 nucleotide stem loop (TR) from the MS2 genome acts as an allosteric effector, guiding conformational switching of the coat protein during capsid assembly. Since the principal conformational changes occur far from the TR binding site, it is important to understand the molecular mechanism underlying this allosteric communication. To this end, we use all-atom simulations with explicit water combined with a path sampling technique to sample the MS2 coat protein conformational transition, in the presence and absence of TR-binding. The calculations find that TR binding strongly alters the transition free energy profile, leading to a switch in the favored conformation. We discuss changes in molecular interactions responsible for this shift. We then identify networks of amino acids with correlated motions to reveal the mechanism by which effects of TR binding span the protein. The analysis predicts amino acids whose substitution by mutagenesis could alter populations of the conformational substates or their transition rates.

I. INTRODUCTION

The controlled interconversion between protein conformational states is crucial for essential cellular functions, including signaling, metabolism, and assembly of the dynamic cytoskeleton. A key regulatory role in such processes is often played by allosteric effectors, whose binding favors a particular protein conformation. The transition pathways by which proteins interconvert between these folded states are largely unknown because intermediates along the pathways cannot be directly characterized by experiments. Similarly, it remains poorly understood how perturbations due to effector binding are communicated across the protein to alter its conformational free energy landscape. In this article we combine long unbiased all-atom molecular dynamics (MD) simulations, an efficient pathway sampling algorithm called the string method¹⁻⁹, and analysis of inter-residue correlations to characterize a protein conformational transition pathway and how it is affected by effector binding. In particular, we study the conformational transition of the MS2 coat protein dimer, and how the binding of an RNA stem loop from the MS2 genome acts as a molecular conformational switch that guides protein assembly into an icosahedrally symmetric capsid.

MS2 is a small bacteriophage that infects male *E. Coli*. During virus assembly, 180 copies of the coat protein (CP) spontaneously assemble around a 3,569 nucleotide single-stranded RNA genome to form an icosahedral capsid. The capsid is a $T=3$ structure, meaning that the CPs adopt three conformations (termed A,B,C) which are precisely arranged within the capsid¹⁰. Major structural differences among the protein conformations are confined to the FG loop, which in the A and C conformations forms an anti-parallel β -hairpin, but in the B conformation is a

flexible loop pulled back against the dimer with a small α -helix kink. The A and C monomers are thus nearly identical, and their FG loops meet at 20 3-fold (quasi-6-fold) axes, whereas the FG loops of the B monomers meet at the 12 5-fold interfaces. In solution, the monomers form stable, non-covalent dimers, which are the basic assembly subunits and will be denoted as CP₂ (Fig. 1 b,c). Formation of the capsid thus requires that 30 CC and 60 AB dimers associate and arrange themselves into the icosahedral geometry (Fig. 1 a).

Based on structural studies, *in vitro* assembly assays, and modeling, it has been proposed that allosteric interactions between CP₂ and the viral genome guide conformational selection during MS2 assembly¹²⁻¹⁵. Capsid assembly can be triggered *in vitro* by the addition of a 19-nucleotide RNA stem-loop (TR) fragment from the genome. TR encompasses the start codon for the replicase protein, and has been shown to bind strongly to the bottom of CP₂^{16,17}. In the crystal structure, TR is bound to the CC dimers in two symmetric orientations, while steric constraints allow only a single orientation for the AB dimer (Fig. 1 g,h).

In vitro experiments by Stockley and coworkers¹² on wild-type CP showed that, in the absence of genomic RNA, CP assembles slowly and produces only a low yield of capsids. Adding a molar ratio of TR results in a strongly bonded CP₂:TR complex that is kinetically trapped. However, adding an equal molar ratio of CP₂ to CP₂:TR results in rapid and efficient assembly. Furthermore, NMR studies on an assembly-incompetent mutant MS2 coat protein (Trp82Arg), showed that TR binding induces a conformation change from a symmetric dimer (presumably BB-like) to an asymmetric dimer (presumably AB-like).

Based on these observations, it was proposed that during assembly of wild-type (WT) MS2 capsid proteins, TR binding acts as a molecular switch which favors a conformational change from the symmetric CC dimer to the asymmetric AB dimer¹². Since both AB and CC dimers are needed for efficient assembly, this scenario is consis-

^{a)}These authors contributed equally

^{b)}Electronic mail: hagan@brandeis.edu

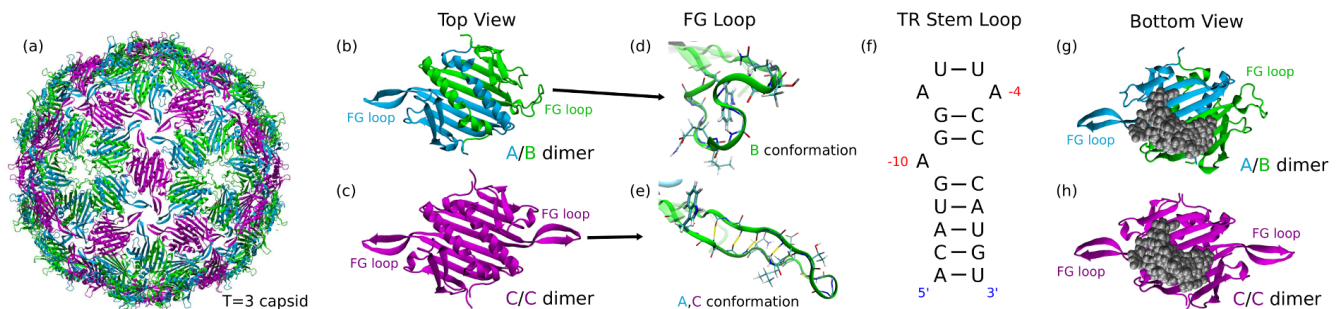


FIG. 1. MS2 Capsid geometry and subunit structure. **(a)** The complete T=3 MS2 capsid of 27.5nm diameter is comprised of 30 CC and 60 AB dimers. It has icosahedral symmetry with the 5-fold vertices as AB dimers and the 3-fold vertices as 3 AB + 3 CC dimers. (pdb ID: 1BMS) **(b)-(c)** the AB and CC dimers colored according to their conformation. The B conformation differs significantly from the A and C conformations in the FG loop. **(d)-(e)** A close up view of the FG loop with a selection of side chains shown as bonds. The B conformation lacks the hydrogen bonds found in the A and C conformations (and shown in yellow). **(f)** The nucleic acid sequence of the TR stem loop, which binds with high affinity to the base of the MS2 dimer. The sequence positions of the adenines that bind most strongly are labeled in red (-10 and -4). **(g)-(h)** MS2 AB and CC dimers shown with the RNA stem loop (TR) bound to their base (pdb ID: 2BU1¹¹). The RNA can adopt two symmetric positions for the CC dimer (only one shown), but the AB dimer allows only one position due to steric collisions. The RNA is shown as grey VDW spheres.

tent with the observation that pure solutions of either CP₂ (assumed to be CC) or CP₂:TR (assumed to be AB:TR) are kinetically trapped whereas an equal molar ratio of CP₂ to CP₂:TR results in rapid and efficient assembly. Subsequent theoretical models suggest that such a conformational switch is consistent with existing structural data and assembly kinetics^{13–15,18}.

Since TR binds CP₂¹⁷ (Fig. 1 g,h) about 12 Å from the FG loop where the conformation change is localized, there is great interest in understanding the molecular mechanism underlying the apparent allosteric communication between these two regions of the protein. Using all-atom normal mode analysis, Dykeman et al.¹³ found that TR binding to an initially symmetric CC conformation leads to asymmetries consistent with the AB conformation. Namely, fluctuations of residues near the FG loop on the A* chain (meaning the chain that corresponds to the A chain in the AB dimer conformation) are suppressed, whereas those near the B* FG loop increase.

The goal of this paper is to directly calculate the MS2 capsid protein conformational free energy landscape, to learn how it is altered by the binding of the genome fragment TR, and to elucidate the molecular basis by which perturbations caused by TR binding are communicated across the protein. To this end, we employ the string method^{1–5} to identify and characterize the most probable transition pathways and associated free energy profiles for the conformational transition in the presence and absence of TR, using all-atom simulations with explicit water. Furthermore, to directly probe the molecular basis for allosteric communication, we characterize correlations of amino acid conformational statistics and motions within long, unbiased MD simulation trajectories. These combined calculations demonstrate that the conformational transition is a complex, multi-step pro-

cess with multiple metastable minima, and is stabilized by multiple molecular-scale interactions whose statistics can be altered by molecular binding in disparate regions of the protein. The analysis predicts several amino acids whose substitution by mutagenesis could alter populations of the conformational substates or their transition rates. These findings may shed light on the mechanisms by which molecular binding affects conformational free energy landscapes in a wide variety of proteins, as well as for understanding the diverse roles of RNA in viral assembly.

Previous computational works have used enhanced sampling methods to examine the effect of small molecule substrates on protein interconversion pathways and free energies, with a particular focus on the enzyme adenylylate kinase^{19–23}. Most closely related to our work, Pattis and May investigated the effect of RNA binding on the Lassa Virus nucleoprotein conformational free energy landscape²⁴.

This article is arranged as follows. In section II we describe the model, simulations, and methodologies used to sample the transition. In section III, we describe the transition pathways predicted by the string method in the presence and absence of TR, we highlight some residues found to play key roles in stabilizing the transition based on the converged strings, and we present results of mutual information on correlations between amino acid conformations and motions. Finally, in section IV we discuss implications of these results for understanding the mechanism underlying the conformational transition and how it is influenced by TR binding. Additional methodological details and validations are given in the appendices.

II. METHODS

A. Systems and simulations

Systems. For statistical analysis and for generating beginning and end points for string method calculations, we initialized unbiased MD simulations from two MS2 capsid protein dimer conformations, each in the presence and absence of the RNA stem loop TR. We denote the four systems as AB, CC, AB:TR, and CC:TR. To avoid complications associated with the fact that P78 undergoes a *cis* to *trans* switch between conformations, we studied P78N mutants, which assemble complete capsids but are not infectious²⁵. The AB and CC dimer structures were therefore extracted from a crystal structure of the empty P78N capsid (pdb ID 1BMS²⁶). Since no crystal structure for P78N capsids with TR is available, we extracted AB:TR and CC:TR from a wild type MS2 capsid containing TR (pdb ID 2BU1), and performed the P78N mutation *in silico* using VMD²⁷. The first and last bases in the RNA stem loop (A and U) for 2BU1 are missing, and were added using CHARMM²⁸.

Each of the four dimer structures was solvated with at least 1nm of water on each side of the structure. The resulting simulation boxes were approximately 10.2nm x 7.7nm x 5.6nm for CP₂ and 10.6nm x 7.2nm x 7.5nm for CP₂:TR. We ensured that each pair of systems intended to serve as beginning and end points of the same string (AB, CC) and (AB:TR, CC:TR) had the same number of atoms. Water molecules were replaced at random with Na⁺ and Cl⁻ ions to neutralize the charge and to bring the ionic strength to 0.1M. The total system size was approximately 41,000 atoms for CP₂ and 58,000 atoms for CP₂:TR. During equilibration, an orientation restraint was added to keep the dimer from self-interaction across the periodic boundary. For long unbiased MD calculations, larger water boxes (approximately 10.2nm³ for both CP₂ and CP₂:TR) were used with no orientation restraints. Details about the equilibration protocol are given in Appendix A 1.

Simulations. Simulations were performed with version 4.5.5 of Gromacs²⁹ modified with version 1.3.0 of the plugin PLUMED³⁰, which was used to generate all restraints and monitor collective variables. Version 5.0.5 of Gromacs was used for the long unbiased simulations. The CHARMM36 all-atom forcefield³¹ was used to represent the system and the TIP3P model^{32,33} was used for water molecules (The string simulations used the CHARMM22/CMAP forcefield^{34,35} for proteins as they were partly performed before the CHARMM36 forcefield was available in Gromacs). Bond lengths were constrained using the LINCS algorithm³⁶ with order 4. The NPT ensemble was simulated using velocity rescaling for the temperature coupling and the Parrinello-Rahman barostat for pressure coupling^{37,38}. Electrostatic interactions were calculated using the particle-mesh Ewald (PME) algorithm³⁹, with a grid spacing of 0.12 and real-space interactions cut off at 1.2nm. Van der Waals inter-

actions were switched at 1.0nm and cut off at 1.2nm.

B. The String Method Algorithm

To determine the minimum free energy transition pathways (MFTP) for the AB \rightleftharpoons CC and AB:TR \rightleftharpoons CC:TR conformations, we used the string method algorithm in collective variables, which was first presented by Maragliano et al⁵. While a number of powerful methods have been developed to sample transition pathways and other rare events (e.g.,^{6,7,40-59}), the string method provides a means to discover the MFTP in a space of many collective variables (CVs), with a computational expenditure that is nearly independent of the number of CVs. To obtain a meaningful free energy minimum, the collective variables must include all slow degrees of freedom relevant to the transition. Our collective variables were chosen to be a subset of the atomic positions^{6,8,60,61}.

The method can be summarized as follows. A set of collective variables capable of characterizing all relevant slow degrees of freedom in the system is chosen (section II C). An initial pathway connecting the two stable states is discretized as an ordered sequence of states (called images) and represented as a curve (called a string) in the multidimensional collective variable space. An iterative calculation is then performed to relax the initial pathway toward a minimum free energy pathway: (1) For each image, multiple short MD simulations are performed in which sampling is constrained to the vicinity of that image in collective variable space by a harmonic bias potential. The gradient of the free energy in collective variable space at each image is calculated from the average force imposed by the bias potential. (2) The position of each image in collective variable space is incremented by displacing it along the (negative) direction of the free energy gradient. (3) Images are redistributed to maintain uniform spacing in arc length along the string. Steps (1-3) are repeated until the string converges to within desired precision. The free energy profile along the converged string can then be calculated using umbrella sampling (section II D).

Details about these procedures, the chosen set of coordinates, and assessments of convergence are given in Appendix B.

C. Selecting Collective Variables

A string is defined by a set of collective variables (CVs), which must include all slow degrees of freedom that are relevant to the reaction. It is not known *a priori* which CVs constitute a good reaction coordinate. While in principle it is possible to choose a large number of CVs with the expectation that a subset of them will constitute a good reaction coordinate, extraneous or redundant CVs can introduce noise that slows convergence. Thus, our goal was to select the minimal possible CV set sufficient

to describe both the CP₂ and CP₂:TR conformational transitions.

Based on extensive trial calculations using various types of CVs (including distances, positions, and dihedral angles), we found Cartesian positions of individual atoms to be best suited for the study of MS2. Atomic positions have been successfully used in previous studies^{6,8,60,61} and can capture both the native CC backbone hydrogen bonds breaking/forming and the formation of the α -kink in the AB state.

Because absolute positions are not invariant under rigid body motions, we restricted translational and rotational diffusion by including position restraints on 10 C α atoms in the top helices of each monomer in CP₂ (residues 105-109). These residues are far from the RNA binding site and FG loop (where the conformational change is localized). In an alternate approach, Ovchinnikov et al. performed principle component analysis on the rigid core of the protein to define a body-centered coordinate system⁶. Another approach is to perform on the fly structural alignment⁶².

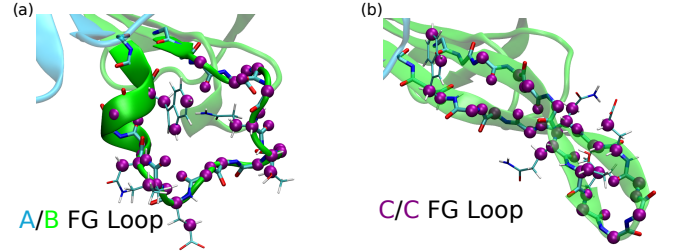
We followed the approach of Ref.⁶ to select the set of atoms whose positions comprise the CVs. We ran a series of targeted molecular dynamics simulations (TMDs)⁶³, in which external biasing forces were applied to the candidate atoms to force the system between conformations. Each candidate set of atoms was ranked by the difference in backbone dihedral angles between the final structure and the target, and the amount of RMSD drift observed during 4ns of simulation after all restraints were released. This test was performed on TMD simulations in both directions (AB to CC and CC to AB) for both CP₂ and CP₂:TR. Once a set of CVs was chosen in this manner, redundant or extraneous atoms were eliminated through a trial and error process in which candidate removals were tested by additional TMD simulations. The final set of CVs contains the positions for 40 atoms, listed in Fig 2.

D. Free Energy Along String Pathway

To calculate the free energy profiles from converged strings, we performed umbrella sampling on an order parameter s that gives the position along the string path. Our implementation is based on the approach described in Ref⁶⁴. To ensure that sampling does not meander arbitrarily far in directions transverse to the transition tube defined by the string⁸, we also defined an order parameter z , which measures the distance from the string. The definition of s and z are inspired by the path collective variables in PLUMED³⁰; for an arc length between images i and $i + 1$, s and z are given by

$$s = i + \frac{(\mathbf{y} - \boldsymbol{\theta}_i) \cdot (\boldsymbol{\theta}_{i+1} - \boldsymbol{\theta}_i)}{|\boldsymbol{\theta}_{i+1} - \boldsymbol{\theta}_i|^2}$$

$$z^2 = \frac{|\mathbf{y} - \boldsymbol{\theta}_i|^2}{|\boldsymbol{\theta}_{i+1} - \boldsymbol{\theta}_i|^2} - (s - i)^2 \quad (1)$$



String CV List: K66:CA, K66:O, V67:CA, A68:N, A68:CA, A68:O, T69:CA, Q70:N, Q70:CA, Q70:CB, Q70:O, T71:CA, T71:CB, V72:N, V72:CA, V72:O, G73:CA, G74:CA, V75:N, V75:CA, V75:O, E76:CA, E76:CG, L77:N, L77:CA, L77:CG, L77:O, N78:N, N78:CA, N78:CB, V79:N, V79:CA, V79:O, A80:CA, A81:N, A81:CA, W82:CA, W82:CB, W82:NE1, W82:CH2

Orientation Restraint Atoms (on both chains): V105:CA, K106:CA, A107:CA, M108:CA, Q109:CA

FIG. 2. The atoms whose positions were selected as CVs for the string are shown for both the AB and CC FG loops. The backbone and a selection of the side chains are shown as bonds, and the string atoms are shown as purple spheres. The string atoms are listed in the text below the figure along with those atoms whose positions were used to prevent translational and rotational diffusion.

where $\boldsymbol{\theta}_i$ gives the vector of CV coordinates defined by image i , \mathbf{y} is the dynamic vector of CV coordinates during sampling, and s is the projection of \mathbf{y} onto the line segment between the bounding images ($\boldsymbol{\theta}_{i+1} - \boldsymbol{\theta}_i$), scaled by the image separation.

During umbrella sampling approximately 150 window centers were spaced evenly in s , with a spring constant of $\kappa \approx 350$ kJ. To maintain sampling near the center of the transition tube⁸, a half-harmonic, upper wall potential was placed between $z = 2$ and $z = 3$, with spring constant $\kappa_{\text{wall}} \approx 450$ kJ.

To check for hysteresis, each window was seeded using steered MD simulations from two different starting structures, one from each of the upper and lower bounding images. The two seeds for each of the ≈ 150 windows were then each sampled for 200ps, so the total simulation time for each free energy calculation was 60 ns. The free energy was calculated from this data using Alan Grossfield's implementation of WHAM (Weighted Histogram Analysis Method)⁶⁵.

E. Root of Mean Squared Fluctuations (RMSF)

The root of mean squared fluctuations (RMSF) for each amino acid about an average structure were calculated for each of the CP₂ and CP₂:TR systems. First, a single 300ns unbiased trajectory was run for each of the four systems and seed structures were extracted as starting points for new trajectories. For the CC, CC:TR and AB:TR system each of the 7 seeded trajectories was 450ns long of which the last 368ns was used in the calculation, resulting in 2.576 μ s of sampling for each system. For the AB system there were 6 seeded trajectories of

530ns length each, from which the last 430ns was used in the calculation, resulting in $2.58\mu\text{s}$ of sampling. Configurations were outputted every 25ps for all the seeded unbiased MD trajectories. The structures from the trajectories were first aligned to minimize the mass-weighted RMSD of the C_α atoms that comprise the core of the protein (residues 7-63 and 85-124 of each monomer). Using the aligned structures, the RMSF was calculated with respect to the average structure and then averaged over all non-hydrogen atoms in each amino acid.

F. Mutual Information

We calculated the mutual information (MI) between all pairs of amino acids for all CP_2 and $\text{CP}_2\text{:TR}$ systems using the approach and MutInf program developed by McClendon et al⁶⁶. In this approach, the MI is calculated using second order terms from the configurational entropy expansion, and indicates the correlation between backbone and side chain conformations⁶⁶. It is calculated using internal coordinates (i.e. the ϕ and ψ backbone dihedrals and side chain rotamers). Amino acids that have shared mutual information have correlated dihedral distributions. Correlated distributions can arise through direct interaction, a chain of interactions, backbone movements, solvent rearrangement, or other mechanisms.

For each system, we applied the MutInf program to the microseconds of multiple trajectories we used in the RMSF calculation. We used 24 bins per degree-of-freedom, and results from 20 sets of scrambled data were calculated and subtracted in order to determine the excess mutual information, as done by McClendon et al⁶⁷. We then used hierarchical clustering on the resulting MI matrix to identify groups of amino acids that share significant mutual information (as done in Ref⁶⁶). We generated the dissimilarity matrix as given in Eq. 2, and used a Euclidean distance metric to cluster amino acids.

$$D_{ij} = \text{Max}(\text{MI}) - \text{MI}_{ij} \quad (2)$$

We systematically extracted the largest possible “real” clusters by recursively splitting the hierarchy of clusters until each cluster achieved an MI average greater than a given cutoff value. After generating the clusters, we verified that they were valid (i.e. had high intra-cluster MI averages and small inter-cluster MI averages) using

$$\text{cluster avg} = \frac{1}{N} \sum_{i \in W_m} \sum_{j \in W_n, j \neq i} \text{MI}_{ij} \quad (3)$$

where W_m is the set of amino acid numbers that belong to cluster m . The sums loop over all residues in W_m and W_n , and N is the total number of elements in the sum such that $i \neq j$. For an intra-cluster average ($m = n$), all amino acid self-correlations are ignored (i.e. $i \neq j$).

From the MI network we calculated the node betweenness centrality for each residue. The node betweenness

centrality determines the number of shortest distance paths that go through each node of the network and is an indicator of the importance of that node in the communication of the network⁶⁸. We used the *tnet* package available through the statistical software R for the calculation of the betweenness centralities⁶⁹.

III. RESULTS

A. Conformational Transition Pathway

In this section we compare the calculated most probable conformational transition pathways for the $\text{CC} \rightleftharpoons \text{AB}$ and $\text{CC:TR} \rightleftharpoons \text{AB:TR}$ MS2 coat protein dimer interconversions. The free energy profiles calculated from the converged strings and illustrative snapshots from the converged strings are shown in Fig. 3. Details on these calculations can be found in Sections II:A,B and Appendix B. Furthermore, several tests of convergence are discussed in Appendix B3. Most significantly, an independent string started from a different initial pathway produced a similar transition pathway and free energy profile (Fig. A3).

1. Pathway in the Absence of TR

The $\text{CC} \rightleftharpoons \text{AB}$ calculation obtains that the symmetric CC state is favored over the AB state by a free energy of $\approx 3k_B T$, and there is one on-pathway metastable state. The string pathway indicates the following order of events for a CC to AB transition with each number corresponding to the snapshots in Fig. 3a,c: The CC loop bends inward, straining the native backbone hydrogen bonds (I \rightarrow II) and eventually breaking them (III), beginning with the bonds closest to the core of the protein (and Trp82). After all of the native hydrogen bonds are broken, the FG loop opens becoming partially solvated and Trp82 leaves its hydrophobic pocket resulting in a metastable state (IV) with a free energy difference of about $12k_B T$ compared to the native CC structure. The FG loop must now widen to accommodate further rotation of Trp82 (V) paying a $\approx 5k_B T$ free energy penalty before collapsing and rearranging into the final AB substate (VI).

2. Pathway in the Presence of RNA

Complexation of CP_2 with the RNA stem loop TR dramatically shifts the free energy landscape, causing the AB:TR substate to be favored over the CC:TR substate by $\approx 20k_B T$. The transition pathway is also markedly different from $\text{CC} \rightleftharpoons \text{AB}$, and now involves two on-pathway metastable states (Fig. 3b,d). In the converged string, the transition from CC:TR to AB:TR proceeds by the following sequence of events with each number corresponding to the snapshots in Fig. 3b,d: The first two backbone hydrogen bonds near the base of the FG

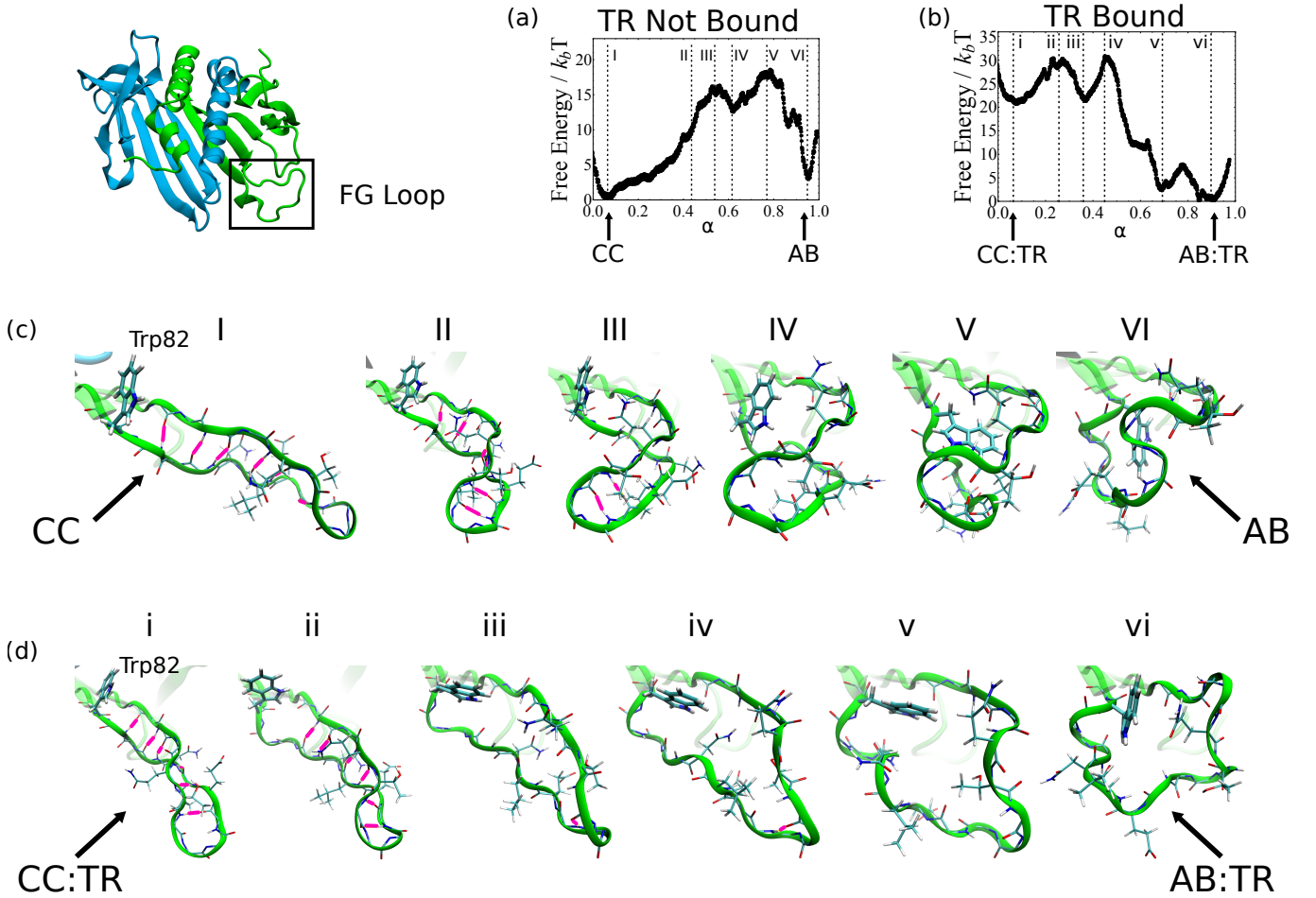


FIG. 3. The most probable transition pathways and associated free energy profiles for $CC \rightleftharpoons AB$ and $CC:TR \rightleftharpoons AB:TR$. (a), (b) The free energy along the most probable pathway as a function of arc length α along the converged strings. (c), (d) Close up snapshots of the B* FG loop along the transition pathway for $CC \rightleftharpoons AB$ (c) and $CC:TR \rightleftharpoons AB:TR$ (d). The native backbone hydrogen bonds of the CC monomer are shown in pink, and side chains with atoms selected as string collective variables (CVs) are shown as bonds. The labels correspond to the position along the free energy profile as indicated in (a) and (b).

loop break (i \rightarrow ii). The backbone dihedral angles of amino acids 79-81 move toward their eventual position in the α -kink of AB:TR. This represents the first barrier to the transition, of $\approx 7k_B T$. It is now free energetically favorable for Trp82 to rotate out of the hydrophobic pocket toward chain A, which forms the first metastable state (iii). Interestingly, this rotation proceeds in the opposite direction as found in the $CC \rightleftharpoons AB$ transition. A free energy penalty of $\approx 7k_B T$ must be paid to reach state iv, which involves side chain rearrangements and further solvation of the FG loop. Finally, Trp82 rotates into the FG loop, which then spontaneously collapses, resulting in the second metastable state (v). This state is structurally very similar to the final AB:TR state (vi), and only $\approx 2k_B T$ higher in free energy. A final rotation of Trp82, involving a $\approx 4k_B T$ barrier, leads to the final AB:TR state (vi).

3. Comparison of Pathways

The most striking difference between the $AB \rightleftharpoons CC$ and $AB:TR \rightleftharpoons CC:TR$ strings is the shift in the most stable sub-state upon TR binding, from CC to AB:TR. This population shift is consistent with experimental data¹². Both pathways highlight the important role of the large side chain of Trp82 in determining the sequence of events during the conformation change. The strings show that the native CC backbone hydrogen bonds require much more substantial molecular rearrangements before breaking and allowing rotation of Trp82 in the $CC \rightleftharpoons AB$ transition as compared to $CC:TR \rightleftharpoons AB:TR$. This difference suggests that the binding of TR destabilizes these hydrogen bonds, which would contribute to shifting the population toward the AB:TR state.

B. Root of Mean Squared Fluctuations

The residue based RMSFs were calculated for the four dimer systems using microseconds of unbiased MD simulations for each system as described in Sec. II E. The RMSF values give the typical fluctuations for each amino acid and provide insight into the relative flexibility of different portions of the protein. These results are summarized in Fig. 4.

Upon TR binding to the CC dimer, the largest change in the RMSF occurs in residues 23-30 of the CD loop, where the fluctuations decrease similarly in the A* and B* monomers (Fig. 4a). Similarly, the loops in both the A* and B* monomers of the RNA bound systems have lower RMSF than those in the non-bound systems (Fig. 4d,e). The CD loop in the CC dimer has higher fluctuations than in any other system. The effect of TR binding on the CD loop can be explained by noting that residues Asn27 and Val29 are in direct contact with TR.

While the CC dimer is symmetric, it is possible that the transition in the B* FG loop to the AB conformation is due in part to asymmetries that arise in dimer fluctuations upon TR binding. The TR induced asymmetry in the fluctuations of the CC:TR system is evident in the RMSF of residues 49 to 53 leading up to and including part of the EF loop, which have higher RMSF in the A* monomer than in the B* monomer (Fig. 4a). This asymmetry is consistent with what was found in a previous all atom normal mode analysis by Dykeman et al.¹³, which showed that for WT MS2 the B factor, which is directly related to the RMSF, of the EF loop decreased in B* and increased in A*. The RNA binding to the AB dimer has the same effect on the EF loop of the B* monomer (Fig. 4c). This effect on the EF loop can be explained by the fact that this part of the protein in the B* monomer is in direct contact with the TR. While Dykeman et al. also found that the B factor of the FG loop in B* increased upon TR binding to the CC dimer, we find the dynamics of the FG loops of the CC:TR system to be similar to the symmetric dimer. However, the RMSF of residues Trp82, Arg83, Tyr85 and Leu86 following the FG loop are suppressed in the A* monomer (Fig. 4a).

A significant difference between the CC:TR and AB:TR systems is in the FG loop, which has higher RMSF in the AB:TR system for both the A* and B* monomers (Fig. 4b,c). Another significant difference between the AB:TR and other systems is in the GH-loop of the B* monomers, where residues 95-100 have higher RMSF in B* monomer of AB:TR 4c,f. The final effect to note is that TR binding in the AB system leads to a dramatic decrease in the dynamic fluctuations of the FG loop of the B* monomer, where in the AB:TR system the dynamics of the loop is much more confined (Fig. 4f). It is important to note that these differences in fluctuations in the B* monomer take place far from the TR binding residues and are a result of long range allosteric communications.

To validate that the sampling is sufficient for this anal-

ysis we present the RMSF of the EF and FG loops of the CC:TR dimer using half of the total trajectory frames in the Appendix C, and show that the results are the same as those using all the frames. In the next section we look at how the changes in the dynamics of TR binding residues are communicated to the rest of the protein, including the FG loop.

C. Mutual Information

Cluster analysis to determine groups of correlated residues. To characterize residue conformational correlations, we also calculated the mutual information between the dihedral angles of all pairs of amino acids from the long unbiased trajectories. We then clustered residues based on their pairwise mutual information to identify groups of residues that are strongly correlated. The intra-cluster averages used as cutoffs for each system were: $0.035k_B T$ for CC, $0.04k_B T$ for AB, $0.07k_B T$ for CC:TR and $0.1k_B T$ for AB:TR. For further discussion on the cutoffs and the resulting number of clusters see Appendix D. To assess sampling convergence, we compared the raw MI data and calculated clusters from all 7 trajectories ($2.576\mu s$) to results obtained using only 4 trajectories ($1.472\mu s$). The Pearson's correlation coefficient for the MI calculated from the full and partial data sets is 0.896, and clusters obtained from the partial data set (Appendix D) are consistent with those from the larger data set.

The clusters from the full data set are presented in Fig. 5 for each of the dimer systems. In the CC and AB dimers, there is one large cluster that encompasses the majority of residues on the TR binding face of the protein (blue in Fig. 5), including both the A* and B* FG loops. Small clusters of residues form at the opposite face of the protein, each primarily containing few residues from the helical domains.

TR binding changes the clustering, particularly of the FG loops. In CC:TR both the A* and B* FG loops are clustered separately from the main body of the protein and regions in direct contact with TR. Trp82, which was found by the string calculations to be crucial for the conformational transition, is also clustered separately from the main protein body in both A* and B* conformations. This implies that the B* FG loop is relatively independent from the TR-bound residues in CC:TR. However, in AB:TR the B* FG loop is clustered with the main body of the protein that is in contact with TR. Hence, the conformation of the B* FG loop in AB:TR is modulated by the bound TR. Recall that the RMSF results showed that the B* FG loop in AB:TR is much more confined than in the AB system. These two results together strongly suggest that the B* FG loop's dynamics is modulated by the TR in the AB:TR conformation. The AB:TR system also stands out as it has higher total MI in the system (Fig. 6a) compared to the other three dimers. Hence, not only does the clustering and RMSF

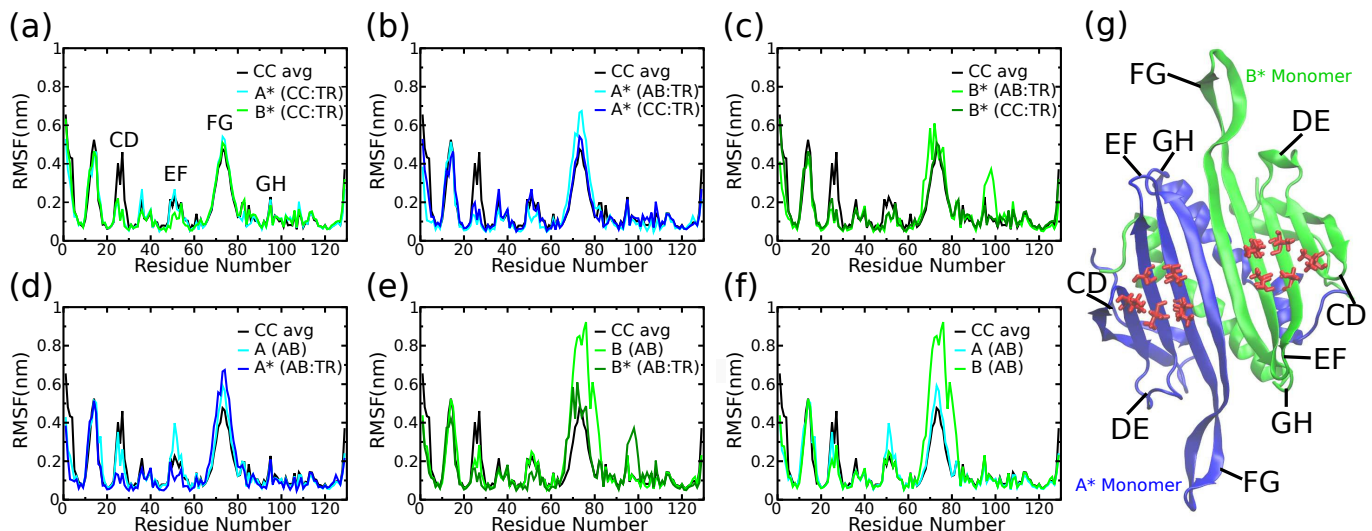


FIG. 4. The RMSF for each residue (averaged over all non-hydrogen atoms within a residue) for each CP₂ and CP₂:TR system. (a)-(f) The RMSF as a function of residue number with the important loops and turns labeled in (a). To facilitate comparison of the dynamics between different systems, each plot shows as a reference the RMSF for the CC dimer, averaged over the two symmetric monomers, as well as the two additional systems. (g) A view of the CC dimer with the loops and turns labeled according to convention⁷⁰. The residues Val29, Thr45, Ser47, Thr59, Lys61 which form the TR binding pocket on each monomer are shown in red.

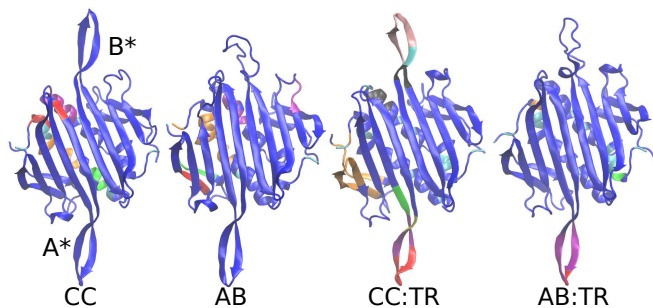


FIG. 5. Mutual information clusters for all the CP₂ and CP₂:TR systems. The clusters of correlated residues, calculated from the mutual information between pairs of amino acids as described in the text, are shown in different colors.

data show that TR binding plays a key role in modulating the B* FG loop through residue correlations in AB:TR, but also that the overall intra-protein communication is stronger in AB:TR than in the other three systems.

Betweenness centrality identifies communication pathways. To gain a molecular-scale understanding of how residue conformational information is transferred across the protein, we filtered the complete mutual informational data to include only pairs of residues in direct contact, defined as residues that are 5.5Å or closer for 75% of the simulation time. We then used this contact-filtered MI graph to calculate the betweenness centrality for each residue as discussed in Sec. II F. Betweenness centrality represents the number of short-

est paths between all node pairs on the MI graph that pass through a given residue, and is thus a measure of how important each residue is for information flow through the network. The betweenness centrality has been used in previous work to determine the importance of nodes in networks constructed from residue-residue interaction energies⁷¹. Here, the network is constructed using residue-residue correlations. In Fig. 6b we show the 11 residues with the highest betweenness centrality ($> 10^4$) for the AB:TR system. We find that there is a group of consecutive residues with high centrality on the B* G β -strand (Tyr85, Leu86, Asn87 and Met88), including Tyr85 which has highest centrality in the system (Fig. 6c). Two other residues with high centrality are Val64 on the B* F β -strand, with the second highest centrality, and Lys66 in the B* FG loop. These results thus suggest a communication network in the protein in which the majority of communication travels through the spine of four residues on the B* G β -strand and is then coupled to the B* F β -strand and FG loop. Perturbations due to TR binding passed along this pathway are thus strongly coupled to the FG loop conformation. In contrast, the same group of residues does not exhibit high centrality in the A* monomer, providing an explanation for why TR binding does not affect the A* FG loop conformation.

Considering the residues with high centrality ($> 10^4$) in the other three dimer systems gives further insight into how intra-protein communication is modulated by TR binding (Fig. 7). For the CC dimer, there is a group of 3 residues (Glu63, Tyr85, and Leu86), that appear in both the A* and B* monomers and are outlined in

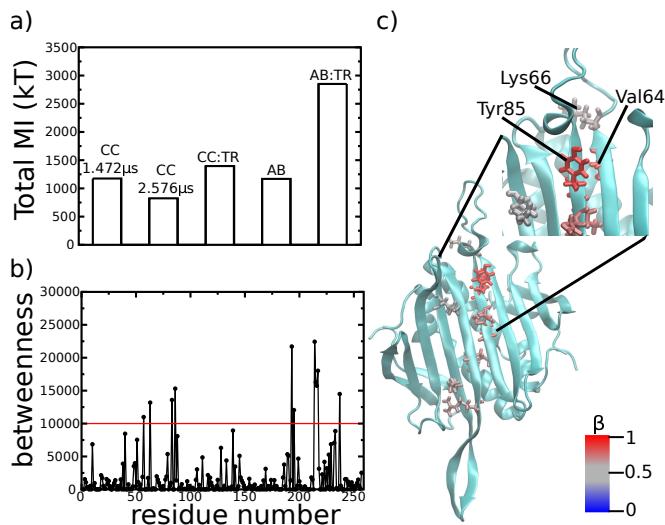


FIG. 6. Total MI of each CP_2 and $CP_2:TR$ system and the betweenness centralities of the AB:TR residues. (a) The total MI of each system is compared. Calculation of the total MI of the CC system for two different sampling sizes gives an estimate of the spread in total MI values due to sampling size variations. The AB:TR system is found to have stronger correlations than the other three systems. (b) The betweenness centralities of each residue in the AB:TR system. Residue numbering starts at the A^* monomer and ends at the B^* monomer. (c) The residues with betweenness centralities greater than 10^4 are represented on a structure of the AB:TR system in stick representation. The color scale corresponds to the relative betweenness β , which is the ratio of the given residue's betweenness centrality divided by the maximum residue betweenness centrality in the system. The residues with betweenness centralities greater than 10^4 are Lys57, Glu63, Arg83, and Leu86 in the A^* monomer and Val64, Lys66, Tyr85, Leu86, Asn87, Met88, and Met108 in the B^* monomer.

Fig. 7. These residues highlight the symmetry that exists in the CC dimer. However, the highest centrality residues are not completely symmetric between the A^* and B^* monomer. For example, residue Met88 appears in the B^* monomer but not in the A^* monomer. We expect these asymmetries to be resolved with greater sampling, but this result also highlights the fact that the instantaneous configurations of the CC dimer are not perfectly symmetric due to fluctuations. The CC:TR results show that TR binding breaks the symmetry between the A^* and B^* monomers present in the CC dimer, since the high centrality residues present in the B^* monomer are not present in the A^* monomer. However, there is a similarity between the B^* monomers in CC and CC:TR, as residues Tyr85, Leu86 and Met88 in the G β -strand appear in both dimers. Tyr85, Leu86 and Met88 also appear in both the A^* and B^* monomers of the AB dimer and the B^* monomer of AB:TR. Hence, these three residues are important for communication in the B^* monomer in all of the dimer systems.

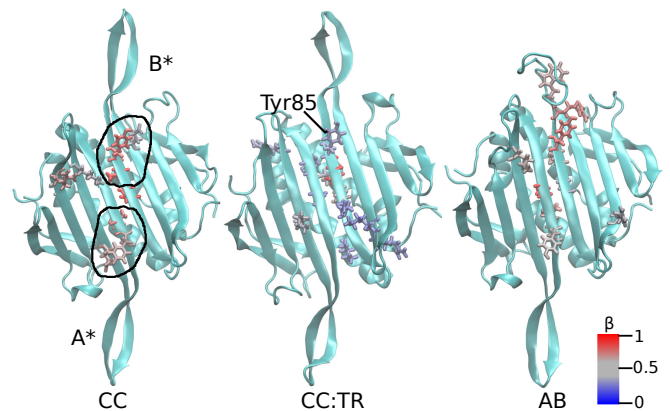


FIG. 7. The relative betweenness β , as described in Fig. 6, shown for residues with a betweenness centrality greater than 10^4 for the CC, CC:TR and AB systems. The three residues Glu63, Tyr85 and Leu86 that appear on both the A^* and B^* monomers of the symmetric CC dimer are outlined. Tyr85 which appears as a high betweenness residue in the B^* monomer of all of the CP_2 and $CP_2:TR$ systems is marked on the CC:TR structure.

IV. DISCUSSION AND CONCLUSIONS

We have combined the string method, free energy calculations, and analysis of long unbiased molecular dynamics simulations to characterize the effect of binding of the MS2 genome fragment TR to its capsid protein. The calculations demonstrate that the impact of TR binding is substantial and far-reaching. The free energy profiles calculated from our converged strings for the $CC \rightleftharpoons AB$ and $CC:TR \rightleftharpoons AB:TR$ transitions (Fig. 3) show a strong shift in the favored population from CC to AB:TR. Furthermore, the strings indicate that TR binding dramatically alters the interconversion pathway, changing the sequence of events and the nature and number of intermediate metastable states. Given that TR binds more than a nanometer from the residues which undergo the majority of conformational rearrangement (the FG loop), our calculations provide direct evidence for allostery and begin to reveal its underlying mechanisms, albeit within the limitations of force field accuracy and finite sampling.

The fundamental effect of TR-binding is to generate an inherently asymmetric dimer. The $CC \rightarrow AB$ transition requires a spontaneous fluctuation that breaks the CC symmetry and ‘chooses’ which FG loop will interconvert to a B conformation. In contrast, TR-binding introduces subunit-spanning asymmetries that favor transition of one chain. We characterized these asymmetries, and how they are transmitted across the protein, by analyzing collective motions and correlated conformational statistics of amino acids within long unbiased MD trajectories of each stable substate. We found extensive asymmetries in both the dynamical fluctuations and correlations. The most significant effect of TR binding was found to be on the FG loop of the B^* monomer when comparing the AB

and AB:TR conformations. We find a pathway of strong communication along a spine of residues between the TR binding region and the B* FG loop, thus identifying how the conformational landscape can be so strongly shifted upon TR binding to stabilize the AB conformation.

Comparison to previous results. Previous experiments on MS2 have shown that TR binding induces a conformation change from a symmetric to an asymmetric structure¹². Based on this and other evidence it has been inferred that the CC state is preferred in the absence of TR, and the AB state in the presence of TR. Our results from the string method calculation and the associated free energy profile directly support this conclusion, and also reveal that the associated transition pathways differ in the presence of TR.

A study by Dykeman et al¹³ performed an all-atom normal mode analysis to determine how the vibrational modes are modified by RNA binding. They found that TR binding to a (WT) CC conformation causes asymmetric fluctuations of the EF loop; fluctuations increase in A* and decrease in B*. The mutant Trp82Arg has an asymmetry in the DE loop instead, which was proposed as a possible explanation for why it is assembly-incompetent. Our MSF calculations also find asymmetries upon TR binding in the EF loop. While there is little difference between the FG loop fluctuations in the CC and CC:TR systems, TR binding in the AB systems shows the fluctuations in the B* FG loop greatly decreased upon TR binding.

Limitations of our calculations and outlook. The relevance of the string method pathway and associated free energy profile depend on the extent to which the collective variables describe all relevant slow degrees of freedom. Furthermore, recent computational studies have shown that conformational transitions can proceed by multiple, diverse pathways (*e.g.*⁶⁴), while a single string calculation typically samples only one transition tube. We have assessed several metrics to determine whether our set of collective variables was sufficient and the extent of sampling within trajectory space: (1) Convergence during string iterations was reasonably rapid. Missing slow degrees of freedom can be expected to slow convergence since they relax on a slow time scale. (2) The umbrella sampling calculations were performed in both directions along the pathway. A lack of significant hysteresis between these two directions is consistent with inclusion of all relevant degrees of freedom. (3) Independent string calculations started from substantially different initial pathways led to very similar converged strings (Fig. A3), consistent with efficient sampling in trajectory space. Taken together, these results are consistent with a sufficient set of collective variables and broad sampling within trajectory space. Further testing of the validity of these calculations could be achieved by comparing the results to those of an independent technique, such as Markov State model (MSM) calculations. It is also possible to use the converged string as a starting point for efficient construction of an MSM⁶⁴.

While we analyzed correlations of amino acid conformations within the stable conformational substates, further insight into the transition mechanism might be obtained by characterizing mutual information during the conformational transition. Finally, in this work, we focused on the effects of TR binding on the coat protein dimer conformations. A natural next step is to examine the effect of TR binding on dimer-dimer interactions; *e.g.* Ref.¹⁴.

ACKNOWLEDGMENTS

This work was supported by the NIH through Award Number R01GM108021 from the National Institute Of General Medical Sciences (MRP and MFH) and R01GM100966 from NIGMS (DTM and MFH). Computational resources were provided by the NSF through XSEDE computing resources (Trestles, Kraken, Queen Bee, and Maverick) and the Brandeis HPCC which is partially supported by the Brandeis Center for Bioinspired Soft Materials, an NSF MRSEC, DMR-1420382.

Appendix A: Additional methods details

1. Equilibration

Each of the 4 systems was relaxed from its initial configuration as follows. First, the system was minimized while iteratively relaxing harmonic restraints on all protein heavy atoms, centered on crystal structure positions. Next, MD simulations were performed in which the same restraints, now centered on the final minimized position, were slowly relaxed as the temperature was gradually increased from 25K to 300K. Unbiased MD was then performed for 50-100ns to ensure equilibration. To prevent self-interaction, rotational drift was limited by harmonic restraints on the α -carbons of residues 105-109 in the A subunit α -helix, which is located on the top of the dimer far from the FG-loop and RNA binding sites.

Of the four systems, only the AB dimer undergoes significant rearrangement. Trp82 in the FG loop rotated and the rest of the FG loop moved toward the EF hairpin of the A monomer; in contrast, the AB:TR FG loop remained close to the DE loop of the B monomer. The difference between equilibrated AB and AB:TR states is significant because Trp82 is a large side chain which rearranges substantially during the conformational changes.

Appendix B: String method calculations

This section presents details on the string method calculation. Following Ovchinnikov et al.⁶, we define N_{cv} collective variables that depend on the Cartesian positions \mathbf{x} of atoms in the protein as $\hat{\theta}(\mathbf{x}) =$

$(\hat{\theta}_1(\mathbf{x}), \hat{\theta}_2(\mathbf{x}), \dots, \hat{\theta}_{N_{cv}}(\mathbf{x}))$. Each image n of the string evolves according to⁶

$$\boldsymbol{\theta}_n(t + \Delta t) = \boldsymbol{\theta}_n(t) - \gamma^{-1} \Delta t \mathbf{M}(\boldsymbol{\theta}_n(t)) \nabla G(\boldsymbol{\theta}_n(t)) \quad (\text{B1})$$

where $\boldsymbol{\theta}_n(t)$ gives the collective variable values of image n from string iteration t and γ is a tuneable “friction constant” that sets the size of the step taken down the free energy gradient (along with Δt). The metric tensor $\mathbf{M}(\boldsymbol{\theta}(t))$ accounts for the curvilinear nature of the collective variables and is given by⁶

$$M_{ij}(\boldsymbol{\theta}) = \sum_k \frac{1}{m_k} \left\langle \frac{\partial \hat{\theta}_i(\mathbf{x})}{\partial x_k} \frac{\partial \hat{\theta}_j(\mathbf{x})}{\partial x_k} \right\rangle_{\hat{\boldsymbol{\theta}}(x)=\boldsymbol{\theta}} \quad (\text{B2})$$

where the sum ranges over each coordinate k for all atoms in the system, $\langle \dots \rangle$ denotes an average over sampling constrained in the vicinity of $\boldsymbol{\theta}$, and m_k is the mass of atom k .

We made two simplifying approximations in our implementation. Since we used only Cartesian coordinates for collective variables, we approximated the tensor $M(\boldsymbol{\theta}_n(t))$ in Eq. (B1) as the identity matrix. Tests with and without this approximation supported that the metric tensor can be neglected for our system.

The second approximation was to dynamically set $\gamma^{-1} \Delta t$ (from Eq. (B1)) such that the step size is a fixed fraction of the image spacing. This guarantees that new images will not jump too far in any given iteration. With the alanine dipeptide model system, we extensively tested our implementation with both approximations against a string implementation with collective variables based on dihedral angles.

To identify collective variables sufficient to describe the transition between states, we systematically vetted candidate coordinates using restrained targeted molecular dynamics simulations⁶ (described for our systems in Appendix II C). Next, we used TMD to generate an initial string connecting the two metastable states. This pathway was then discretized into images, and the string was systematically relaxed by the following iterative procedure.

1. **Sample.** For each image n , run short simulations to estimate $\nabla G(\boldsymbol{\theta}_n)$ (the free energy gradient in collective variable space in the proximity of image n). In each short simulation, impose a harmonic potential for each collective variable, centered on the image. The spring constant of the harmonic potential is selected to keep sampling in the vicinity of its image (typically with an average sampling radius of 1-2 image spacings). Calculate the average force imposed by each potential.
2. **Evolve.** Generate a new string by displacing each image a distance δ in the direction opposite to the free energy gradient. In our implementation, δ is scaled to be a fixed fraction of the image spacing.

3. **Reparameterize.** Redefine the locations of images along the string so that they are uniformly spaced in arc length, following the implementation of Maragliano et al.⁵.

This procedure was iterated until the string pathway approximately converged, which was assessed by the RMSD between the initial and current strings. We define the RMSD between strings as

$$\text{RMSD}(\boldsymbol{\theta}^{S_1}, \boldsymbol{\theta}^{S_2}) = \left(\int_0^1 |\boldsymbol{\theta}^{S_1}(\alpha) - \boldsymbol{\theta}^{S_2}(\alpha)|^2 d\alpha \right)^{1/2} \quad (\text{B3})$$

with $\boldsymbol{\theta}^{S_i}(\alpha)$ as the N_{cv} -dimensional point at fraction α along string S_i . Since the strings are discretized, it is necessary to interpolate between images.

1. Generating the Initial String

Initial strings were generated from TMD trajectories of the CC→AB and CC:TR→AB:TR transitions, with the TMD bias based only on CV atoms. Coordinates were saved every 2ps and used to construct a time series of CV values, which was then smoothed to prevent noise from dominating image selection. The data was smoothed by applying a nearest neighbor smoothing kernel to the coordinates for 10-20 iterations. Forty images, with approximately equal spacing, were then selected from the smoothed trajectory and used as the initial string pathway. The spacing between the N_{cv} -dimensional images in the initial string was 3.5Å for CP₂ and 2.5Å for CP₂:TR, which provided sufficient resolution to capture bond-breaking and all significant conformational rearrangements.

TMD parameters. During selection of collective variables and generation of initial pathways, TMD simulations imposed a harmonic potential as a function of the RMSD difference between the current and target structure, measured from the positions of the candidate CV atoms only. The center of the potential was moved linearly from the RMSD of the initial configuration to 0 over 1.5 ns. The spring constant was linearly scaled from $k = 2.5 \times 10^6 \text{ kJ/mol} \cdot \text{nm}^2$ to $k = 5 \times 10^6 \text{ kJ/mol} \cdot \text{nm}^2$ over this same interval. After centering on RMSD=0, k was linearly increased over three separate 500ps intervals to $k = (2, 20, 200) \times 10^7 \text{ kJ/mol} \cdot \text{nm}^2$. After this, k was linearly decreased to 0 over 1ns, followed by 4ns of unbiased simulation.

2. Running the String

Each string was evolved according to the three steps outlined at the beginning of Appendix B: sample, evolve, reparameterize. In the sample step, for each image n , the structure from the previous (or initial) string closest in CV space to the image CV values $\boldsymbol{\theta}_n$ was sub-

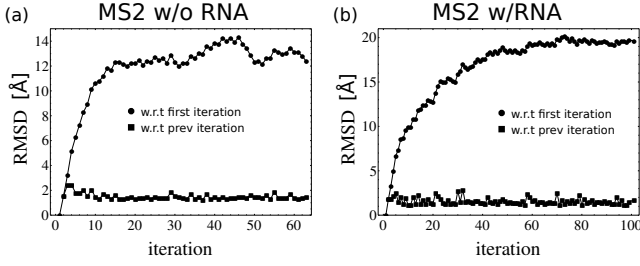


FIG. A1. The RMSD during the equilibration of the (a) CP_2 and (b) $CP_2:TR$ strings. The RMSD is calculated according to Eq. (B3) with respect to both the initial string and the previous string iteration.

jected to a steered molecular dynamics (SMD) simulation targeting θ_n . A harmonic potential with force constant $k_{\text{drag}} = 10^5 \text{ kJ/nm}^2$ was imposed for each CV, and moved linearly to θ_n (at a speed of no faster than $0.1 \text{ nm}/(1000 \text{ steps})$). Then, we sampled the local free energy gradient by performing MD for an additional 200ps with harmonic restraints for the CV centered at θ_n . To maintain local sampling while speeding convergence, we chose a restraint force constant of $k_{\text{hold}} = 450.0 \text{ kJ/nm}^2$, yielding an average sampling radius of 1-2 image spacings. CV values were recorded every 0.1ps. The string was then evolved by updating CV values according to Eq. B1, with a step size set to a fixed fraction $\delta = 0.5$ of the image spacing, followed by reparameterization to maintain uniform spacing along the arc length.

To monitor string convergence, we calculated the RMSD (in CV space) between points at equal arc length along the string according to Eq. B3. We used a linear interpolation between neighboring images to calculate the RMSD at arc lengths not commensurate with image locations. Each string was run until the RMSD with respect to the initial string plateaued, which required 50-100 iterations (Fig. A1).

3. String Convergence and Validity

To test the assumption that a plateau of the RMSD in CV space is a good measure of string convergence, we calculated the $CC \rightleftharpoons AB$ free energy profile for two string iterations after the RMSD plateau (Fig. A2). Although the two free energy profiles are not identical, they obtain the same free energy difference between CC and AB substates, nearly the same barrier height, and both have a single on-pathway metastable state.

To assess global convergence of the string, we performed a second string calculation for the $CC:TR \rightleftharpoons AB:TR$ transition, initialized from a different TMD simulation. This TMD used a slightly different definition for the CC:TR and AB:TR substates, and produced an initial pathway which differs substantially from the initial pathway used for the first string.

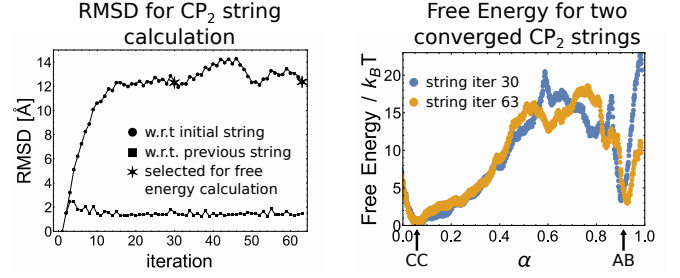


FIG. A2. (a) The string RMSD as calculated from Eq. (B3) during the convergence of a CP_2 string. The strings at iterations 30 and 63 were taken for the free energy calculation (as marked by the stars). (b) The free energy profile for each of the two converged CP_2 strings as a function of arc length α .

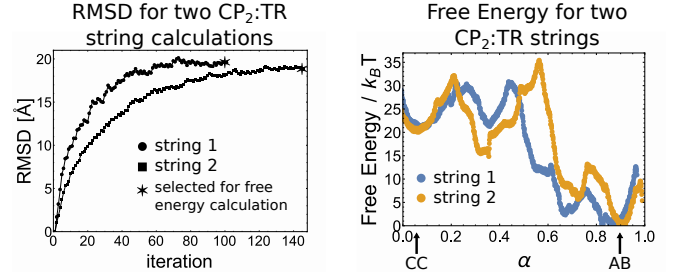


FIG. A3. (a) The RMSD with respect to the initial string (Eq. B3) during the convergence of $CP_2:TR$ strings initialized from two independent TMDs. (b) The free energy profiles for the two final $CP_2:TR$ strings as a function of arc length α .

The RMSD of 13 \AA between the two initial pathways is approximately as large as the RMSD between the first converged string and its initial pathway. The convergence and resulting free energy profiles are shown in Fig. A3. Once again, the two strings result in the same relative free energies for the CC:TR and AB:TR substates and contain the same number (two) of on-pathway metastable states. While there are quantitative differences, the overall similarity between the two calculations suggests that the strings have converged to the same pathway. This result from two different initial pathways is consistent with a global MFTP, although a thorough assessment would require a number of additional strings and hence a large computational cost.

Appendix C: RMSF

To validate that the sampling used is sufficient for the RMSF calculations we compare the per residue RMSF for the EF and FG loop regions for the CC:TR system using the full data set of $2.576 \mu\text{s}$ and using half the total trajectory frames. These results are presented in Fig. A4 and are marked “full” and “half”. The RMSF values

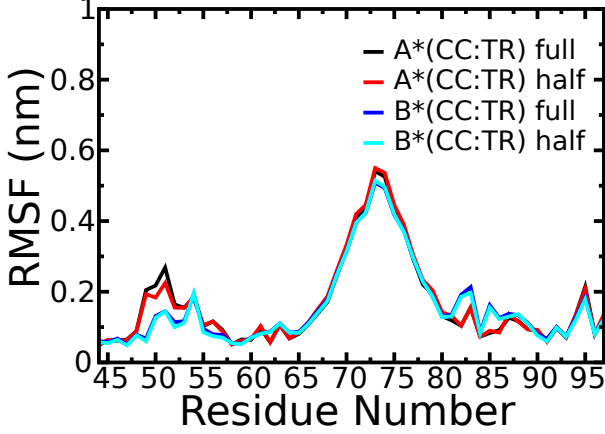


FIG. A4. The per residue RMSFs of the EF and FG loop regions of the A* and B* monomers of the CC:TR dimer. The results using the full simulation data and half the simulation data are compared and show that the results obtained using the full simulation data are converged.

obtained from using the full data set are the same as what was presented in Fig. 4a. The results show that the two data sets give very similar results and that the calculated RMSF values are converged.

Appendix D: Mutual Information Calculations

To determine groups of amino acids that have correlated distributions, we used hierarchical clustering on the MI matrices. In hierarchical clustering, each amino acid starts in its own cluster, and clusters with minimal “dissimilarity” are recursively merged until only one cluster remains. For our calculation, the “dissimilarity” was determined by the intra-cluster average of D_{ij} of Eq. (2). From the resulting hierarchy of clusters, we systematically extracted the largest possible clusters, such that the intra-cluster MI average was greater than a certain cutoff value. Hierarchical clustering was applied to each of the four MI data sets. An intra-cluster average cutoff of $0.035k_B T$ was used for the CC system, which results in 7 distinct clusters. Increasing the cutoff to $0.04k_B T$ lead to a jump in the number of clusters to 12, hence we decided to base our analysis on the smaller 7-cluster result. A similar approach was used for the other three dimer systems, where a cutoff was found to generate a reasonable number of clusters with low inter-cluster MI averages, but stayed below an intra-cluster average cutoff that would lead to a jump in the number of clusters. Hence, an intra-cluster average cutoff of $0.04k_B T$ was used for the AB system, which results in 5 distinct clusters. Increasing the cutoff to $0.05k_B T$ lead to an increase in the number of clusters to 10. For the CC:TR sys-

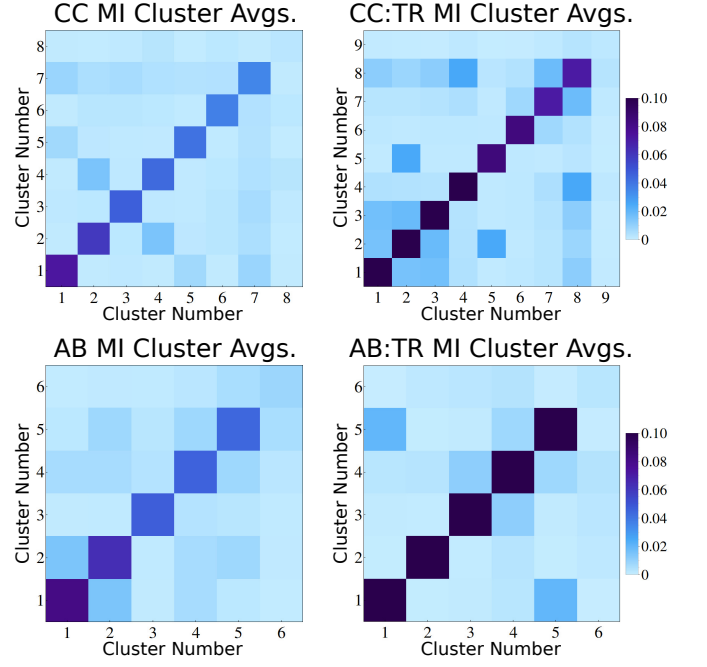


FIG. A5. The similarity matrix for the mutual information matrix clusters for each CP₂ and CP₂:TR system. The coloring indicates the average mutual information and is scaled uniformly for all plots (as shown on the right in units of $k_B T$). The last cluster in each plot contains amino acids that do not share MI with each other or other clusters, which is why the last diagonal element has a value near 0.

tem an intra-cluster average cutoff of $0.07k_B T$ was used, resulting in 8 distinct clusters. Increasing the cutoff to $0.08k_B T$ lead to an increase in the number of clusters to 14. For the AB:TR system an intra-cluster average cutoff of $0.10k_B T$ was used, resulting in 5 distinct clusters. Increasing the cutoff to $0.11k_B T$ lead to an increase in the number of clusters to 10.

The resulting clusters presented in Fig. 5 were tested to ensure that they have a high intra-cluster correlation average and a low inter-cluster correlation average (as calculated from Eq. 3). The resulting correlations are shown in Fig. A5, where the diagonal shows strong intra-cluster correlations.

The MI matrix for the CC dimer was clustered using all of the 7 trajectories ($2.576\mu s$) as shown in Fig. 5 and using 4 trajectories ($1.472\mu s$). The resulting clusters for both sampled MI data sets for CC are compared in Fig. A6. Using an intra-cluster average cutoff of $0.05k_B T$ lead to 8 distinct clusters. The cluster identifications are similar in for the $2.576\mu s$ and $1.472\mu s$ sampled systems. Importantly, the FG loops in the A* and B* monomers are clustered with the large blue cluster that includes the TR binding sites. Hence, we conclude that the sampling is converged for the calculation of the MI clusters.

¹E. Weinan, W. Ren, and E. Vanden-Eijnden, Phys. Rev. B **66** (2002).

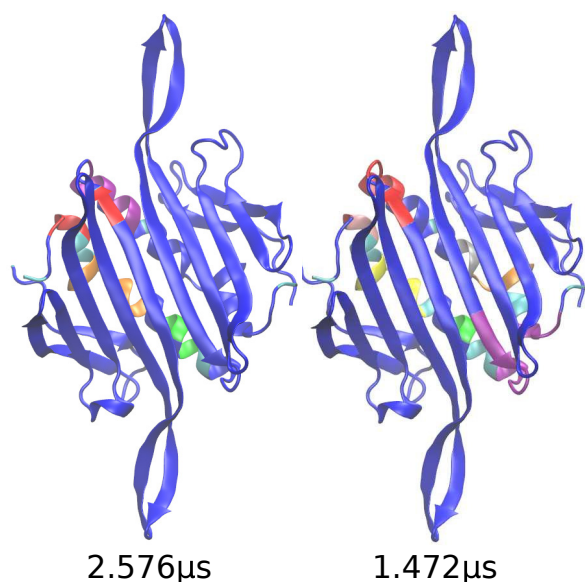


FIG. A6. The clusters of the MI matrix for the CC system using the $2.576\mu\text{s}$ and $1.472\mu\text{s}$ of sampling. The similarity in the resulting clusters indicates that the sampling is converged.

- ²W. E. Ren, and E. Vanden-Eijnden, *J. Chem. Phys.* **126** (2007).
- ³W. E. and E. Vanden-Eijnden, *Annu Rev Phys Chem* **61**, 391 (2010).
- ⁴E. Weinan, W. Ren, and E. Vanden-Eijnden, *Chem. Phys. Lett.* **413**, 242 (2005).
- ⁵L. Maragliano, A. Fischer, E. Vanden-Eijnden, and G. Ciccotti, *J. Chem. Phys.* **125** (2006).
- ⁶V. Ovchinnikov, M. Karplus, and E. Vanden-Eijnden, *J. Chem. Phys.* **134** (2011).
- ⁷A. C. Pan, D. Sezer, and B. Roux, *J. Phys. Chem. B* **112**, 3432 (2008).
- ⁸V. Ovchinnikov and M. Karplus, *J. Chem. Phys.* **140**, 175103 (2014).
- ⁹L. Maragliano, B. Roux, and E. Vanden-Eijnden, *J. Chem. Theory Comput.* **10**, 524 (2014).
- ¹⁰D. L. D. Caspar and A. Klug, *Cold Spring Harbor Symp. Quant. Biol.* **27**, 1 (1962).
- ¹¹E. Grahn, T. Moss, C. Helgstrand, K. Fridborg, M. Sundaram, K. Tars, H. Lago, N. J. Stonehouse, D. R. Davis, P. G. Stockley, and L. Liljas, *RNA* **7**, 1616 (2001).
- ¹²P. G. Stockley, O. Rolfsson, G. S. Thompson, G. Basnak, S. Francese, N. J. Stonehouse, S. W. Homans, and A. E. Ashcroft, *J. Mol. Biol.* **369**, 541 (2007).
- ¹³E. C. Dykeman and R. Twarock, *Phys. Rev. E* **81**, 031908 (2010).
- ¹⁴K. M. ElSawy, L. S. D. Caves, and R. Twarock, *J. Mol. Biol.* **400**, 935 (2010).
- ¹⁵V. L. Morton, E. C. Dykeman, N. J. Stonehouse, A. E. Ashcroft, R. Twarock, and P. G. Stockley, *J. Mol. Biol.* **401**, 298 (2010).
- ¹⁶J. Gott, T. Pan, K. Lecuyer, and O. Uhlenbeck, *Biochemistry* **32**, 13399 (1993).
- ¹⁷P. G. Stockley, N. J. Stonehouse, J. B. Murry, S. T. S. Goodman, S. J. Talbot, C. J. Adams, L. Liljas, and K. Valegard, *Nucleic Acids Res.* **23**, 2512 (1995).
- ¹⁸E. C. Dykeman, N. E. Grayson, K. Toropova, N. A. Ranson, P. G. Stockley, and R. Twarock, *J. Mol. Biol.* **408**, 399 (2011).
- ¹⁹S. L. Seyler and O. Beckstein, *Mol. Simulat.* **40**, 855 (2014).
- ²⁰K. Arora and C. L. Brooks, *Proc. Natl. Acad. Sci. USA* **104**, 18496 (2007).
- ²¹S. Lu, W. Huang, and J. Zhang, *Drug Discovery Today* **19**, 1595 (2014).
- ²²Y. Matsunaga, H. Fujisaki, T. Terada, T. Furuta, K. Moritsugu, and A. Kidera, *PLoS Comput. Biol.* **8**, e1002555 (2012).
- ²³Y. Wang, A. D. Hollingsworth, S. K. Yang, S. Patel, D. J. Pine, and M. Weck, *J. Am. Chem. Soc.* **135**, 14064 (2013).
- ²⁴J. G. Pattis and E. R. May, *Biophys. J.* **110**, 1246 (2016).
- ²⁵H. Hill, N. J. Stonehouse, S. A. Fonseca, and P. G. Stockley, *J. Mol. Biol.* **266**, 1 (1997).
- ²⁶N. J. Stonehouse, K. Valegård, R. Golmohammadi, S. van den Worm, C. Walton, P. G. Stockley, and L. Liljas, *J. Mol. Biol.* **256**, 330 (1996).
- ²⁷W. Humphrey, A. Dalke, and K. Schulten, *J. Mol. Graph.* **14**, 33 (1996).
- ²⁸B. R. Brooks, C. L. Brooks, III, A. D. Mackerell, Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caffisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoseck, W. Im, K. Kucsera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, *J. Comput. Chem.* (2009).
- ²⁹D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, *J. Comp. Chem.* **26**, 1701 (2005).
- ³⁰M. Bonomi, D. Branduardi, G. Bussi, C. Camilloni, D. Provasi, P. Raiteri, D. Donadio, F. Marinelli, F. Pietrucci, and R. a. Broglia, *Comp. Phys. Comm.* **180**, 1961 (2009).
- ³¹R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig, and A. D. MacKerell, *J. Chem. Theory Comput.* **8**, 3257 (2012).
- ³²W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, *J. Chem. Phys.* **79**, 926 (1983).
- ³³D. J. Price and C. L. Brooks, *J. Chem. Phys.* **121**, 10096 (2004).
- ³⁴A. D. J. MacKerell, D. Bashford, M. Bellott, R. L. J. Dunbrack, J. D. Evans, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kucsera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. I. Reiher, B. Roux, M. Schlenkerich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wierkiewicz-Kucsera, D. Yin, and M. Karplus, *J. Phys. Chem. B* **102**, 3586 (1998).
- ³⁵A. D. Mackerell, *J. Comput. Chem.* **25**, 1584 (2004).
- ³⁶B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, *J. Comput. Chem.* **18**, 1463 (1997).
- ³⁷M. Parrinello and A. Rahman, *J. Appl. Phys.* **52**, 7182 (1981).
- ³⁸S. Nosé and M. L. Klein, *Mol. Phys.* **50**, 1055 (1983).
- ³⁹U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, *J. Chem. Phys.* **103**, 8577 (1995).
- ⁴⁰M. A. Rohrdanz, W. Zheng, and C. Clementi, in *Annu. Rev. Phys. Chem.*, Annual Review of Physical Chemistry, Vol. 64 (Annual Reviews, 2013) pp. 295–316.
- ⁴¹A. Dickson and A. R. Dinner, *Annu. Rev. Phys. Chem.* **61**, 441 (2010).
- ⁴²P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, *Annu. Rev. Phys. Chem.* **53**, 291 (2002).
- ⁴³S. Fischer, K. W. Olsen, K. Nam, and M. Karplus, *Proc. Natl. Acad. Sci. USA* **108**, 5608 (2011).
- ⁴⁴R. Elber, *Biophys. J.* **92**, L85 (2007).
- ⁴⁵F. Pietrucci, F. Marinelli, P. Carloni, and A. Laio, *J. Am. Chem. Soc.* **131**, 11811 (2009).
- ⁴⁶M. Lei, M. I. Zavodszky, L. A. Kuhn, and M. F. Thorpe, *J. Comput. Chem.* **25**, 1133 (2004).
- ⁴⁷D. Moroni, P. G. Bolhuis, and T. S. van Erp, *J. Chem. Phys.* **120**, 4055 (2004).
- ⁴⁸R. J. Allen, P. B. Warren, and P. R. ten Wolde, *Phys. Rev. Lett.* **94** (2005).
- ⁴⁹J. Pfaffentner, D. Branduardi, M. Parrinello, T. D. Pollard, and G. A. Voth, *Proc. Natl. Acad. Sci. USA* **106**, 12723 (2009).
- ⁵⁰A. Barducci, M. Bonomi, and M. Parrinello, *Biophys. J.* **98**, L44 (2010).
- ⁵¹B. W. Zhang, D. Jasnow, and D. M. Zuckerman, *Proc. Natl. Acad. Sci. USA* **104**, 18043 (2007).
- ⁵²G. A. Huber and S. Kim, *Biophys. J.* **70**, 97 (1996).

- ⁵³A. L. Ferguson, A. Z. Panagiotopoulos, I. G. Kevrekidis, and P. G. Debenedetti, *Chem. Phys. Lett.* **509**, 1 (2011).
- ⁵⁴T. S. Van Erp, “Dynamical rare event simulation techniques for equilibrium and nonequilibrium systems,” (John Wiley & Sons, Inc., 2012) pp. 27–60.
- ⁵⁵G. R. Bowman, V. A. Voelz, and V. S. Pande, *J. Am. Chem. Soc.* **133**, 664 (2011).
- ⁵⁶F. Noe, C. Schutte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, *Proc. Natl. Acad. Sci. USA* **106**, 19011 (2009).
- ⁵⁷C. Chen, *J. Phys. Chem. B* **120**, 3061 (2016).
- ⁵⁸R. Elber, *J. Chem. Phys.* **144**, 060901 (2016).
- ⁵⁹H. Fujisaki, K. Moritsugu, Y. Matsunaga, T. Morishita, and L. Maragliano, *Front. Bioeng. Biotechnol.* **3**, 125 (2015).
- ⁶⁰J. J. Lacroix, S. A. Pless, L. Maragliano, F. V. Campos, J. D. Galpin, C. A. Ahern, B. Roux, and F. Bezanilla, *J. Gen. Physiol.* **140**, 635 (2012).
- ⁶¹C. Zhao and S. Y. Noskov, *Plos Comput. Biol.* **9**, e1003296 (2013).
- ⁶²D. Branduardi and J. D. Faraldo-Gomez, *J. Chem. Theory Comput.* **9**, 4140 (2013).
- ⁶³J. Schlitter, M. Engels, and P. Krger, *J. Mol. Graph.* **12**, 84 (1994).
- ⁶⁴F. Pontiggia, D. Pachov, M. Clarkson, J. Villali, M. Hagan, V. S. Pande, and D. Kern, *Nat. Commun.* **6**, 8284 (2015).
- ⁶⁵A. Grossfield, “Wham: the weighted histogram analysis method, version 2.0.9,”.
- ⁶⁶C. L. McClendon, G. Friedland, D. L. Mobley, and M. P. Jacobson, *J. Chem. Theory Comput.* **5**, 2486 (2009).
- ⁶⁷C. L. McClendon, A. P. Kornev, M. K. Gilson, and S. S. Taylor, *Proc. Natl. Acad. Sci. USA* **111**, E4623 (2014).
- ⁶⁸S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, *Physics Reports* **424**, 175 (2006).
- ⁶⁹T. Opsahl, F. Agneessens, and J. Skvoretz, *Soc. Networks* **32**, 245 (2010).
- ⁷⁰R. Golmohammadi, K. Fridborg, M. Bundule, K. Valegrd, and L. Liljas, *Structure* **4**, 543 (1996).
- ⁷¹A. A. S. T. Ribeiro and V. Ortiz, *Biophys. J.* **109**, 1110 (2015).